Exploration de Données Temporelle

session pratique accélérée
avec Matlab

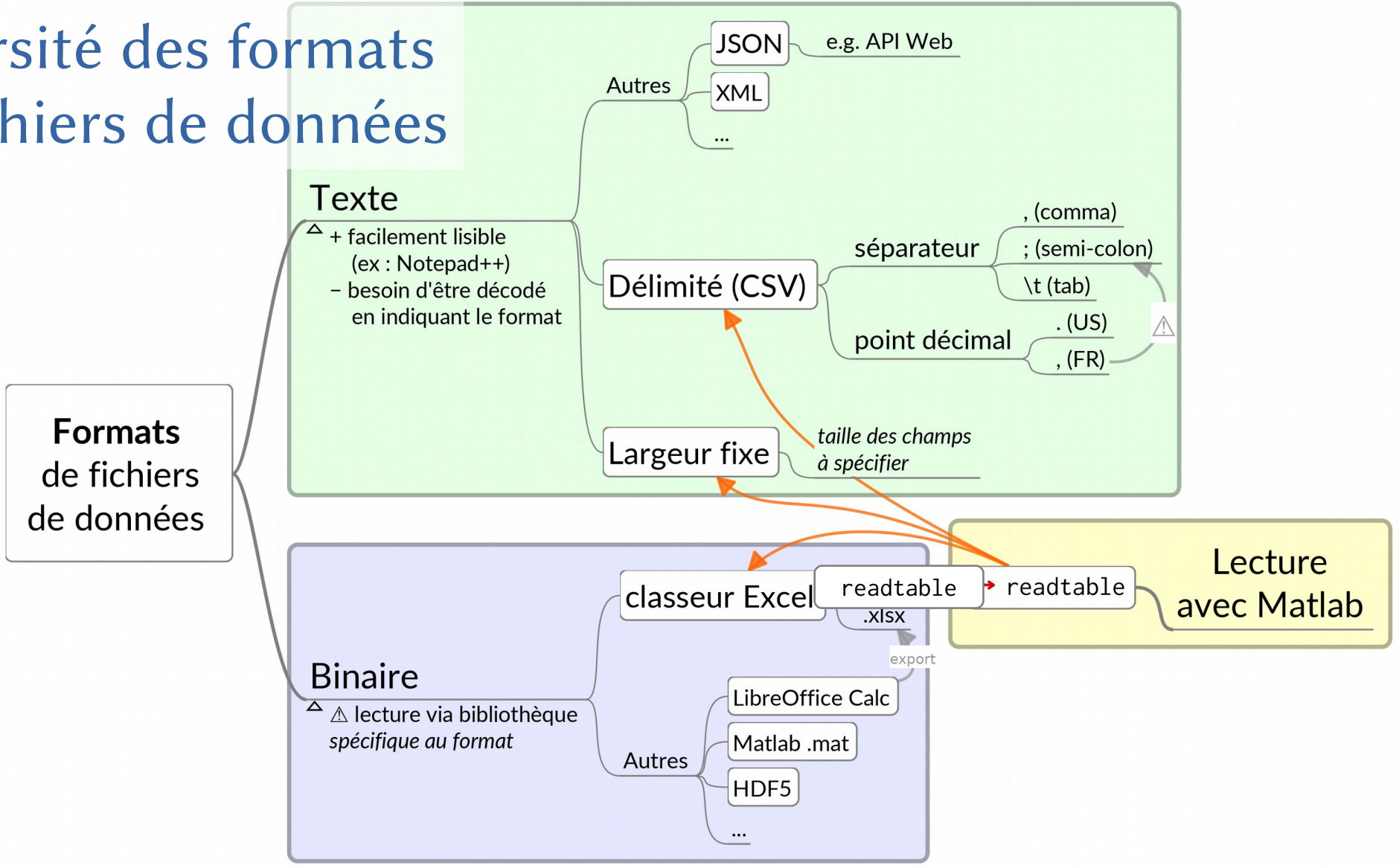
Pierre Haessig, Marie-Anne Lefebvre
CentraleSupélec, Rennes, Octobre 2019

Plan de la séance

- Introduction : qq notions
- Chargement & prétraitements des données
  - Diversité des formats de fichiers de données
  - Données manquantes ou aberrantes
- Mise en forme & synchronisation de données

Introduction : qq notions

Diversité des formats de fichiers de données



Exemple de formats texte

- CSV variante A

```
2019-10-01,15.1,80
2019-10-02,12.1,50
2019-10-03,8.0,10
```

- Largeur fixe

```
201910011580
201910021250
201910030810
```

NB : padding de zéros

- CSV variante B

```
2019;10;1;15,1;80
2019;10;2;12,1;50
2019;10;3;8,0;10
```

- JSON array

```
[
  {
    "Date": "2019-10-01",
    "Température": 15,
    "Humidité": 80
  },{…},{…}
]
```

Chargement de données avec Matlab

- Fonction `readtable` : Excel, Texte délimité (CSV) et largeur fixe
  - et sa compagne `detectImportOptions`
- Outil graphique “Import Tool” (clic-droit fichier : Importer...)
  - 🙂 pratique au premier abord
  - ☹ mais au final *peu automatisable*
    (sauf l’option pour générer une fonction d’importation)

Outils pour l'analyse de données temporelles avec Matlab

- Type container « `table` »
  - et la version spécialisée « `timetable` » pour les données temporelles
- Plus adapté que les matrices « classiques » de Matlab :
  - **Spécialisation statistique** : Colonne = Variable, Lignes = Observations
  - Possibilité de colonnes hétérogènes (nombres, catégories, texte…)
  - Petites fonctions pratiques : `head`, `tail`, `summary`

⚠ différencier le type du *conteneur* et le type du *contenu* (*verre* vs *eau*)

Intérêts de chaque fichier proposé

	Météo Rennes	Log capteur qualité de l'air	Fiche présence
Date	4 colonnes numériques	2 colonnes texte (date, heure)	1 colonne texte
Structure	Wide	Narrow (empilée)	
Données aberrantes	Oui		
Données catégorielles		Oui	Oui
Échantillonnage	Régulier	Irrégulier	Irrégulier

Pratique : météo à Rennes

Archive de la station météo Rennes St-Jacques :

- Fichier : `meteo_rennes_2015.txt`
  - Toute l'année 2015
- Source : NOAA Integrated Surface Database (ISD) https://www.ncdc.noaa.gov/isd

Pratique : météo à Rennes

Objectif :

- Charger les données
- Extraire la température
- Caractériser et tracer la température,
  - sur l'année
  - sur une période donnée

Enjeux :

- Types de fichiers de données
- Données manquantes/aberrantes
- Données de type `datetime`

Valeurs manquantes & aberrantes

- Doc : Missing Data in MATLAB
- Les caractériser (détection, comptage)
- Les marquer

Pratique : Log capteur qualité de l'air

Log capteur de qualité de l'air (T°, CO_2...) installé en salle 404

- Fichier : `log-20150309-171821.csv`

NB : la salle 404 est la salle de réunion de l'équipe d'Automatique du campus de Rennes.

Pratique : log capteur qualité de l'air

Objectif :

- Charger les données
- **Resynchroniser** les mesures des différents sous-capteurs

Enjeux :

- Désempilement (format Wide vs Narrow)
- Décodage (*parsing*) des dates au format texte
  - et opération inverse : représentation textuelle de dates dans la console
- Rééchantillonage, agrégation

Format de tableau : Wide vs Narrow

Person	Age	Weight	Height
Bob	32	128	180
Alice	24	86	175
Steve	64	95	165

Person	Variable	Value
Bob	Age	32
Bob	Weight	128
Bob	Height	180
Alice	Age	24
Alice	Weight	86
Alice	Height	175
Steve	Age	64
Steve	Weight	95
Steve	Height	165

https://en.wikipedia.org/wiki/Wide_and_narrow_data

Variables catégorielles

- Wikipedia : Variable catégorielle, Matlab : categorical,
- Utilisation statistique : Grouping Variables → boxplot, grpstats, gscatter...
- Enjeux d'efficacité :
  - Stockage en mémoire
  - Traitements (ex. : comparaisons)

Catégorie
"Catégorie A"
"Catégorie B"
"Catégorie A"
"Catégorie A"
"..."

Catégorie	Table de correspondance
1	1: "Catégorie A"
2	2: "Catégorie B"
1	
1	
... (1 octet)	

Formats de date texte

- Date « yyyy-MM-dd » ou 3 colonnes Y,M,D
  - Mais aussi « dd/MM/yyyy » (FR) ou « MM/dd/yyyy » (US) ‼
- Heure « HH:mm:ss »
- Timestamp UNIX (https://www.epochconverter.com/ http://timestamp.fr/)
- Et on ne parlera pas des fuseaux horaires et des heure été/hiver… (sauf pour conseiller de tout enregistrer en UTC). Et encore moins de des secondes intercalaires.

Échantillonage et Synchronisation

Pratique : Fiche de présence

Nombres de personnes présentes en salle 404
(noté au moment d'un changement)

- Fichier : `fiche_presence.csv`

Pratique : Fiche de présence

Objectif :

- Charger les données et les synchroniser avec les donnée de qualité de l'air
- Corréler la concentration en CO_2 avec la présence

Enjeux :

- « Recap final »
- Resynchronisation